EURASIP Journal on Advances in Signal Processing
a SpringerOpen Journal

RESEARCH Open Access

Multi-pose lipreading and audio-visual speech recognition

Virginia Estellers* and Jean-Philippe Thiran

Abstract

In this article, we study the adaptation of visual and audio-visual speech recognition systems to non-ideal visual conditions. We focus on overcoming the effects of a changing pose of the speaker, a problem encountered in natural situations where the speaker moves freely and does not keep a frontal pose with relation to the camera. To handle these situations, we introduce a pose normalization block in a standard system and generate virtual frontal views from non-frontal images. The proposed method is inspired by pose-invariant face recognition and relies on linear regression to find an approximate mapping between images from different poses. We integrate the proposed pose normalization block at different stages of the speech recognition system and quantify the loss of performance related to pose changes and pose normalization techniques. In audio-visual experiments we also analyze the integration of the audio and visual streams. We show that an audio-visual system should account for non-frontal poses and normalization techniques in terms of the weight assigned to the visual stream in the classifier.

1 Introduction

The performance of automatic speech recognition (ASR) systems degrades heavily in the presence of noise, compromising their use in real world scenarios. In these circumstances, ASR systems can benefit from the use of other sources of information complementary to the audio signal and yet related to speech. Visual speech constitutes such a source of information. Mimicking human lipreading, visual ASR systems are designed to recognize speech from images and videos of the speaker's mouth. This fact gives rise to audio-visual automatic speech recognition (AV-ASR), combining the audio and visual modalities of speech to improve the performance of audio-only ASR, especially in presence of noise [1,2]. In these situations, we cannot trust the corrupted audio signal and must rely on the visual modality of speech to guide recognition. The major challenges that AV-ASR has to face are, therefore, the definition of reliable visual features for speech recognition and the integration of the audio and visual cues when taking decisions about the speech classes.

A general framework for AV-ASR [1,3] has been developed during the last years, but for a practical deployment the systems still lack robustness against non-ideal working conditions. Research has particularly neglected the variability of the visual modality subject to real scenarios, i.e., non-uniform lighting and non-frontal poses caused by natural movements of the speaker. The first studies on AV-ASR with realistic conditions [4,5] applied directly the systems developed for ideal visual conditions, obtaining poor lipreading performance and failing to exploit the visual modality in the multi-modal systems. These studies pointed out the necessity of new visual feature extraction methods robust to illumination and pose changes. In particular, the topic of pose-invariant AV-ASR is central for the future deployment of this technology in genuine AV-ASR applications, e.g., smart-rooms or in-car vehicle systems. In these scenarios the audio modality is degraded by noise and the inclusion of visual cues can improve recognition. However, in natural situations the speaker moves freely, a frontal view to the camera is rarely kept and pose-invariant AV-ASR is necessary. It can be considered, then, as the first step in the adaptation of laboratory AV-ASR systems to the conditions expected in real applications.

In lipreading systems, the variations of the mouth's appearance caused by different poses are more significant than those caused by different speech classes and, therefore, recognition degrades dramatically when non-

* Correspondence: virginia.estellers@epfl.ch
Signal Processing Laboratory LTS5, Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

frontal poses are matched against frontal visual models. It is necessary to develop an effective framework for pose invariant lipreading. In particular, we are interested in pose-invariant methods which can easily be incorporated in the AV-ASR systems developed so far for ideal frontal conditions and reduce the train/test mismatch derived from pose changes. Techniques to adapt ASR systems to working conditions have already been developed for the audio modality (Cepstral mean subtraction [6] and RASTA processing [7]), but equivalent methods are necessary for the visual modality. In fact, the same problem exists in face recognition and several methods proposed for pose-invariant face recognition [8-11] can be applied to the lipreading problem. Motivated by these studies and the potential of AV-ASR in human-computer interfaces [12], we propose to introduce a pose normalization step in a system designed for frontal views, i.e., we generate virtual frontal views from the non-frontal images and rely on the existing frontal visual models to recognize speech. The pose normalization block has also an effect on the audio-visual fusion strategy, where the weight associated to the visual stream in the speech classifier should reflect its reliability. We can expect the virtual frontal features generated by pose normalization to be less reliable than the features extracted directly from frontal images. Therefore, the weight assigned to the visual stream on the audio-visual classifier should also account for the pose normalization.

Previous study on this topic is limited to Lucey et al. [13,14], who projected the final visual speech features of complete profile images to a frontal viewpoint with a linear transform. However, the authors do not justify the use of a linear transform between the visual speech features of different poses, are limited to the extreme cases of completely frontal and profile views and their audiovisual experiments are not conclusive. Compared to these studies, we introduce other projection techniques applied in face recognition to the lipreading task and discuss and justify their use in the different feature spaces involved in the lipreading system: the images themselves, a smooth and compact representation of the images in the frequency domain or the final features used in the classifier. We also analyze the effects of pose normalization in the audio-visual fusion strategy in terms of the weight associated to the visual stream. Lucey et al. [13] propose an audio-visual system based on the concatenation of audio and visual features in a single stream, which is later processed in the speech classifier neglecting the multi-modal nature of speech and the possibility to assign different weights to the audio and visual streams. The main contributions of this study, partially presented in [15], are the adaptation of pose-invariant methods used in face recognition to the lipreading system, the study of linear regression for pose normalization in different feature spaces and the study of its effects on the weight associated to the visual stream in the classifier. Our experiments are the first comprehensive experimental validation of pose normalization in visual and audio-visual speech recognition, analyzing the adaptation of laboratory AV-ASR systems to the conditions expected in real applications.

The article is organized as follows. First, we review the structure of an AV-ASR system and explains how the pose-invariance is introduced. We then present the techniques adopted in face recognition to obtain a multi-pose system, adapt some of them to the lipreading problem and study the different feature spaces where the pose normalization can take place. Finally, we report experimental results for visual and audio-visual ASR systems and present the conclusions of our study.

2 Audio-visual speech recognition

In terms of the visual modality, AV-ASR systems differ in three major aspects: the visual front-end, the audio-visual integration strategy and the pattern classifier associated to the speech recognition task. In Figure 1, we present a typical AVSR system. First, the audio front-end extracts the audio features that will be used in the classifier. This block is identical to that of an audio-only ASR system and the features most commonly used are perceptual linear predictive [16] or Mel frequency cepstral coefficients [17,18]. In parallel, the face of the speaker has to be localized from the video sequence and the region of the mouth detected and normalized before relevant features can be extracted [1,19]. Typically, both audio and visual features are extended to include some temporal information of the speech process. Then, the features are used in statistical classifiers, usually hidden Markov models (HMM) [20], to estimate the most likely sequence of phonemes or words. The fusion of information between modalities can happen at two stages [1]: merging the extracted features before going through pattern classifiers or on the statistical models used for classification. In the following, we focus on the visual modality, in particular in the blocks affected by the pose changes on the speaker: the extraction of visual features from images of the mouth and the integration of the visual and audio streams. Finally, we describe the standard AV-ASR system that we adopt and describe how pose normalization can be included in it.

2.1 Visual front-end

The first task on the visual front-end is to identify and extract a normalized region of interest (ROI), which is usually a rectangle centered on the mouth of the speaker [1,21,22]. The normalization of the ROI requires a robust method to detect the face and extract centered,

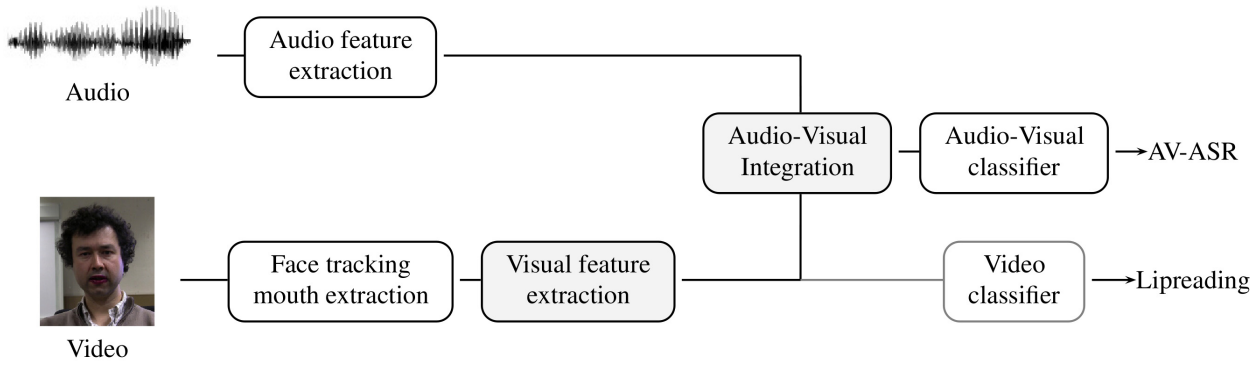

Figure 1 Standard AV-ASR system. Structure of audio-visual ASR system. Upper row corresponds to the audio system, where the features used for speech recognition are extracted and fed to the audio-visual integration block and classifier. The lower part corresponds to the lipreading system: first the mouth is tracked and a sequence of normalized mouth images is extracted, then the visual features are computed and finally used in the audio-visual integration and classification blocks.

aligned, and scaled images of the mouth for each sequence to makes recognition invariant to small movements of the speaker [19]. This preprocessing step is not part of the lipreading system and it is usually included in the face detection block because the position of the mouth, its size and alignment are determined in relation to other face features (the eyes, the tip of the nose). However, an accurate extraction of the mouth ROI is critical in lipreading systems and induced the term *front-end effect* to refer to the effects of the ROI extraction in the performance of the speech recognition system. In that sense, the use of markers or special lipstick on the speaker avoids the use of more complicated mouth tracking techniques [19] to alleviate the front-end effect.

Two main types of features are used for visual speech recognition: appearance based features extracted directly from the pixels of the ROI [1,21,22] and shape based features extracted from the contour of the speaker's lips [23]. Several studies [24,25] report that appearance-based features outperform shape based ones and are, therefore, the features commonly chosen in lipreading and AV-ASR systems. In this approach, the pixels of the ROI themselves are used as features and, consequently, locating the ROI needs to be done with very good precision [26] and the front-end effect carefully considered. The dimensionality of the obtained feature-vector is too large to allow an accurate statistical modeling in the classifiers and dimensionality reduction techniques are necessary. The most popular of these techniques are image compressing transforms [27], as principal components analysis [21,22] or the discrete cosine transform (DCT) [1]. They reduce the size of the images by eliminating redundancy, but there is no guarantee that they are appropriate for the classification task. Linear discriminant analysis (LDA) [28] is a transform capturing relevant information for classification and is thus commonly used in AV-ASR. Other supervised transforms based on ideas from information theory have also been proposed for AV-ASR [29-32], but LDA is widely used because it is simple (linear), gives good results and can easily incorporate dynamic information. Dynamic features measure the visual motion during speech and are more robust to skin color or illumination conditions than the original features. This motion can be represented either by delta images or transforms measuring the inter-frame change of the features, e.g., inter-frame LDA [1].

2.2 Audio-visual integration and classification

Audio-visual integration can be grouped into two categories: feature and decision fusion [1,3]. In the first case, the audio and visual features are combined projecting them onto an audio-visual feature space, where traditional single-stream classifiers are used [33-36]. Decision fusion, on its turn, processes the streams separately and, at a certain level, combines the outputs of each single-modality classifier. Decision fusion allows more flexibility for modality integration and is the technique usually adopted [1,3], in AV-ASR systems because it allows to weight the contribution of each modality in the classification task.

In the statistical models used in AV-ASR, the features of the audio and visual streams are assumed class conditionally independent [37,38], the joint probability distribution is then factorized into single-stream distributions and stream weights λ_A, λ_V are introduced to control the importance of each modality in the classification task [39,40]. The resulting joint probability distribution reads

$$p\left(x_A, x_V | q = q_i\right) = p\left(x_A | q = q_i\right)^{\lambda_A} p\left(x_V | q = q_i\right)^{\lambda_V}, \qquad (1)$$

where x_A, x_V are the audio and visual features and q the class variable. This weighting scheme is naturally introduced in the HMM classifiers by means of multi-stream HMMs [41]. In multi-stream HMMs, independent statistical models like Gaussian mixtures [42] are used to compute the likelihood of each stream independently, which are then combined accordingly to the integration technique. In early integration the streams are assumed to be state synchronous and the likelihoods are combined at state level as indicated by Equation (1). Late integration, in its turn, combines the likelihoods at utterance level, while in intermediate integration the combination takes place at intermediate points of the utterance. The weighting scheme, nonetheless, remains the same and early integration is generally adopted [3]. A common restriction is that the weights λ_A, λ_V sum up to one, which assures that the relation between the emission likelihoods and transition probabilities is kept the same as in single-stream HMMs.

2.3 Our lipreading system

Our speech recognition system is similar to the state-of-the-art presented in [1,3], which we take as a model and introduce in it the pose normalization. On the following, we describe our system, giving more details for the blocks which play a role on the pose normalization task.

In order to minimize the front-end effect, we work with sequences where the speaker wears blue lipstick and we can accurately track the mouth by color information. Our work focuses on the adaptation of the visual features for pose normalization and the use of lipstick sequences allows us to decouple the performance of the face tracker (optimized for frontal poses and whose performance depends on the head pose and illumination) from the extraction of accurate visual features, which is critical in the case of appearance-based features. In our sequences the mouth ROI is detected in the hue domain and normalized mouths of 64×64 pixels are extracted for the definition of the visual features.

In the second block of our system, state-of-the-art audio and visual features are extracted. In terms of audio features, we adopt Mel Frequency Cepstral Coefficients (MFCC), together with their first and second time derivatives and their means removed by Cepstral mean subtraction [6]. For the visual counterpart, we choose appearance-based features and the following sequence of dimensionality reduction transforms. From the original ROI images x_I (frontal) and y_I (lateral), we extract a compact low-dimensional representation of the image space retaining only the first 140 DCT coefficients in zig-zag order in x_F, y_F. To normalize the features for different speakers and sequences, we remove their mean value over the sequence in an equivalent technique to the Cepstral mean subtraction applied to the audio features, and finally LDA transforms are applied to further reduce the dimensionality of the features and adapt them to the posterior HMM classifier. First, intra-frame LDA reduces to 40 the dimensionality of the features while retaining information about the speech classes of interest (phonemes). Afterwards, inter-frame LDA incorporates dynamic information by concatenating 5 intra-frame LDA vectors over adjacent frames and projecting them via LDA to the final features x_L, y_L, which have dimension 39 and will be modeled by the HMMs.

The classifiers used are single- and weighted multi-stream HMMs [43]. In the case of AV-ASR, the use of weighted multi-stream HMMs incorporates the audio-visual integration into the classification task, which is done at state level with the weights leading to best performance on an evaluation data.

In our system, see Figure 2, we assume the pose to be known and introduce a pose normalization block to create virtual frontal features from non-frontal ones at different stages of the visual feature extraction: the extracted mouth ROI (x_I, y_I), a smooth and compact representation of the images in the frequency domain (x_F, y_F) or the final LDA features used in the classifier (x_L, y_L). When the transformations are applied directly to the image space, the pose normalization takes place after the mouth extraction, indicated by number 1 in Figure 2. In case of applying the pose-transformation to the selected DCT or LDA features, the transformation block is introduced after the corresponding feature extraction, numbered 2 and 3 in Figure 2.

3 Pose-invariant lipreading

In this section, we present the techniques adopted in face recognition to obtain a multi-pose system, justify the choice of linear regression (LR) as the technique best suited to our AV-ASR system and study the different feature spaces where the pose normalization can take place.

3.1 From face recognition to lipreading

The techniques proposed for pose-invariant face recognition can be classified into viewpoint transform and coefficient-based techniques [8]. Coefficient based techniques estimate the face under all viewpoints given a single view, either by defining pose-invariant features known as "face lightfields" [44] or estimating the parameters of 3-D face models [45]. In the viewpoint transform approach the face recognition system is designed and optimized for the dominant view (frontal) and a preprocessing step transforms the input images corresponding to undesired poses to the desired view [8]. The same two strategies can be applied to the lipreading task. We adopt the viewpoint-transform approach

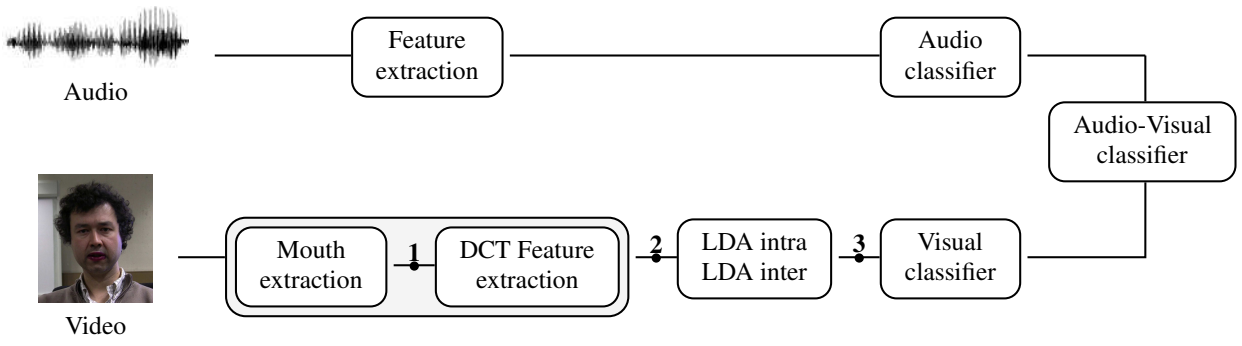

Figure 2 AV-ASR system adopted in our experiments. Block diagram of the AV-ASR system adopted in our experiments. We consider this system as a model and introduce in it a pose normalization block in order to allow multi-pose speech recognition. The lower part corresponds to the lipreading system, where the pose normalization is incorporated after the mouth has been tracked and normalized. The possible feature spaces where the pose normalization can take place are related to the different steps involved in the computation of the visual features: first the images themselves, then a low-dimensional representation in frequency domain and finally the LDA features designed for the classification task. The audio-visual fusion is also adapted in order to take into account the reliability associated to the visual stream after pose normalization.

because lipreading predominantly takes place with frontal views and coefficient-based techniques would suffer from over-generalization [8], i.e., only a small fraction of the time the system would benefit from the definition of pose-invariant features, while most of the time it would be outperformed by a system optimized for frontal views.

In the viewpoint transform approach there are two strategies to generate virtual frontal views from non-frontal poses: 3-D models [9,10] and learning-based methods [46,47]. In the first case, a 3-D morphable model of the face must be built from 2-D images before virtual views from any viewpoint can be generated with graphic rendering techniques. It is computationally expensive and time consuming to match the input 2-D image with the 3-D model and, therefore, that technique is not aimed to the real-world applications of AV-ASR. To overcome that issue, learning-based approaches learn how to estimate virtual views directly in the 2-D domain, either via a 2-D face model or from the images themselves. Working directly with the images, a simple and yet effective way to project the images from lateral to frontal views is based on linear regression [8,11]. Several reasons justify the use of the images themselves instead of introducing a mouth model to estimate the virtual views of the mouth. First, most lipreading systems use directly images of the mouth as visual features and do not require mouth or lip models, which we do not want to introduce only for the pose normalization [3]. Second, the visual features extracted from the images themselves are more informative than features based on lip-modeling, as they include additional information about other speech articulators such as teeth, tongue, and jaws also useful in human speech perception [48]. Besides, appearance based features directly obtained from the image pixels are generic and can be applied to mouths of any viewpoint compared to lip models which have to be developed for any possible view. Finally, these pose normalization techniques involve transforms that can be quickly computed and allow real-time implementations required in most AV-ASR applications.

3.2 Linear regression in multi-pose face recognition

Given a set of M training examples of the undesired viewpoint $Y = [y^1 \dots y^M]$ and their synchronous examples on the target viewpoint $X = [x^1 \dots x^M]$, a matrix W performing LR is determined minimizing the cost function Q

$$Q(W) = \sum_{i=1}^{M} \left\| x^i - W y^i \right\|^2 + \beta \|W\|^2, \tag{2}$$

which measures the mean square error on the training dataset and might include a Tykhonov regularization term (weighted by parameter β) introducing additional smoothness properties and leading to a ridge regression [49]. The well-known solution to the LR is given by $W = XY^T (YY^T + \beta I)^{-1}$, with I the identity matrix.

Linear regression is theoretically justified when images of the same object but from different poses are subject to the same illumination. In the case of face recognition, in [11] Chai et al. show that if the face images are well aligned, there exists an approximate linear mapping $x_I = W^I y_I$ between images of one person captured under variable poses x_I and y_I, which is consistent through different people. Unfortunately, in real-world systems face images are only coarsely aligned, occlusions derived from the 3-D nature of faces affect the different views and the linear assumption no longer holds. To this end,

the authors propose the use of a piecewise linear function to approximate the non-linear mapping existing between images from different poses. The main idea of the proposed method lies in the following intuitive observation: partitioning the whole face into multiple patches reduces the face misalignment and variability between different persons and the transformation associated to pose changes can be better approximated with a linear map for the patches than for the whole image. That technique is called local LR (LLR) in opposition to the previous implementation of LR, which considered the images as a whole and is therefore designated as global LR (GLR).

Intuitively, LLR partitions the whole non-frontal image into multiple patches and applies linear regression to each patch. Given the training set $\{X, Y\}$, each face image is divided into blocks of rectangular patches $\{X_i, Y_i\}_{i = 1...N}$. Then, for each pair of frontal and lateral patches the linear regression matrix W_i is computed as in the GLR case. In the testing stage, any input image with known pose is partitioned into patches, which are used to predict the corresponding frontal patches with the LLR matrices $\{W_i\}_{i = 1,...,N}$. Afterwards, all the virtual frontal patches are combined into a whole vector to construct a virtual frontal image. The patches can be adjacent or overlap, alleviating in that case the block effect but increasing the cost of reconstruction as the value associated to a pixel sampled by several patches is then computed as the mean of the specific pixels in the overlapping patches. Consequently, the patch size and overlapping are parameters to choose for the LLR method to succeed. While a too large patch size suffers from the linear assumption and can lead to blurring of the images, a patch too small is more sensible to misalignments and produces artefacts on the reconstructed image. The overlapping criteria, on its turn, is a trade-off between over-smoothing (high overlapping of patches) and introducing block effects on the reconstructed images (adjacent patches). For the frontal views a uniform partition of the images is adopted, while for non-frontal images each patch contains surface points of the same semantics as those in the corresponding frontal patch. In the case of a completely profile image, for instance, we associate two frontal patches to each profile patch by imposing symmetry on the frontal view. See Figures 3 and 4 for an example of a pair of patches defined across different views.

3.3 Linear regression and lipreading

In our study, the LR techniques are applied considering X and Y to be either directly the images from frontal and lateral views X_I, Y_I or the visual features extracted from them at different stages of the feature extraction process. A first set of features X_F, Y_F are designed to smooth the images and obtain a more compact and low-dimensional representation in the frequency domain. Afterwards, those features are transformed and their dimensionality again reduced in order to contain only information relevant for speech classification, leading to the vectors X_L, Y_L used in the posterior speech classifier.

The visual features X_F, Y_F are the first coefficients of the two-dimensional DCT of the image following the zigzag order, which provide a smooth, compact and low dimensional representation of the mouth. Note that the selected DCT can be obtained as a linear transform, $X_F = SDX_I$, with D the matrix of two dimensional DCT basis transform and S a matrix selecting the DCT coefficients of interest. Therefore there is also an approximate linear mapping $DW^I D^T$ between the DCT coefficients of the frontal and lateral images x_D, y_D. Indeed, as the DCT forms an orthonormal base in the image space, we can write the original linear mapping between the images as

$$x_D = Dx_I = DW^I y_I = DW^I \left(D^T D\right) y_I = DW^I D^T y_D. \quad (3)$$

Consequently, if all DCT coefficients are selected and $S = I$, the DCT coefficients obtained from W^I by projecting images y_I and the W^F projected coefficients from y_F coincide. The linear relationship, however, no longer holds when we consider only a reduced set of DCT coefficients $x_F = SDx_I$ and the transform W^F found with the LR method should be considered an approximation of the non-linear mapping existing between any pair of reduced DCT coefficients x_F and y_F. In that case, selecting the DCT features corresponding to lower frequencies to compute the transform W^F corresponds to smoothing the images previous to the projection and estimating a linear transform forcing the projected virtual image to be smooth by having only low-frequency components. Moreover, the lower-dimensionality of X_F, Y_F compared to X_I, Y_I improves accuracy of the LR matrix estimation due to the *Curse of Dimensionality* [50], which states that the number of samples necessary to estimate a vectorial parameter grows exponentially with its dimensionality. In that sense, the effect of the regularization parameter β is more important in the estimation of W^I than W^F, as imposing smoothness reduces the number of required samples.

It is important to note that the proposed LLR technique on the DCT features provides a different meaning to the patches. If we choose the patches to be adjacent blocks of the DCT coefficients, we are considering different transforms for different frequency components of the image. Consequently, we use an equal partition of the selected DCT coefficients to define the frontal and associated lateral patches in the LLR transform. In that

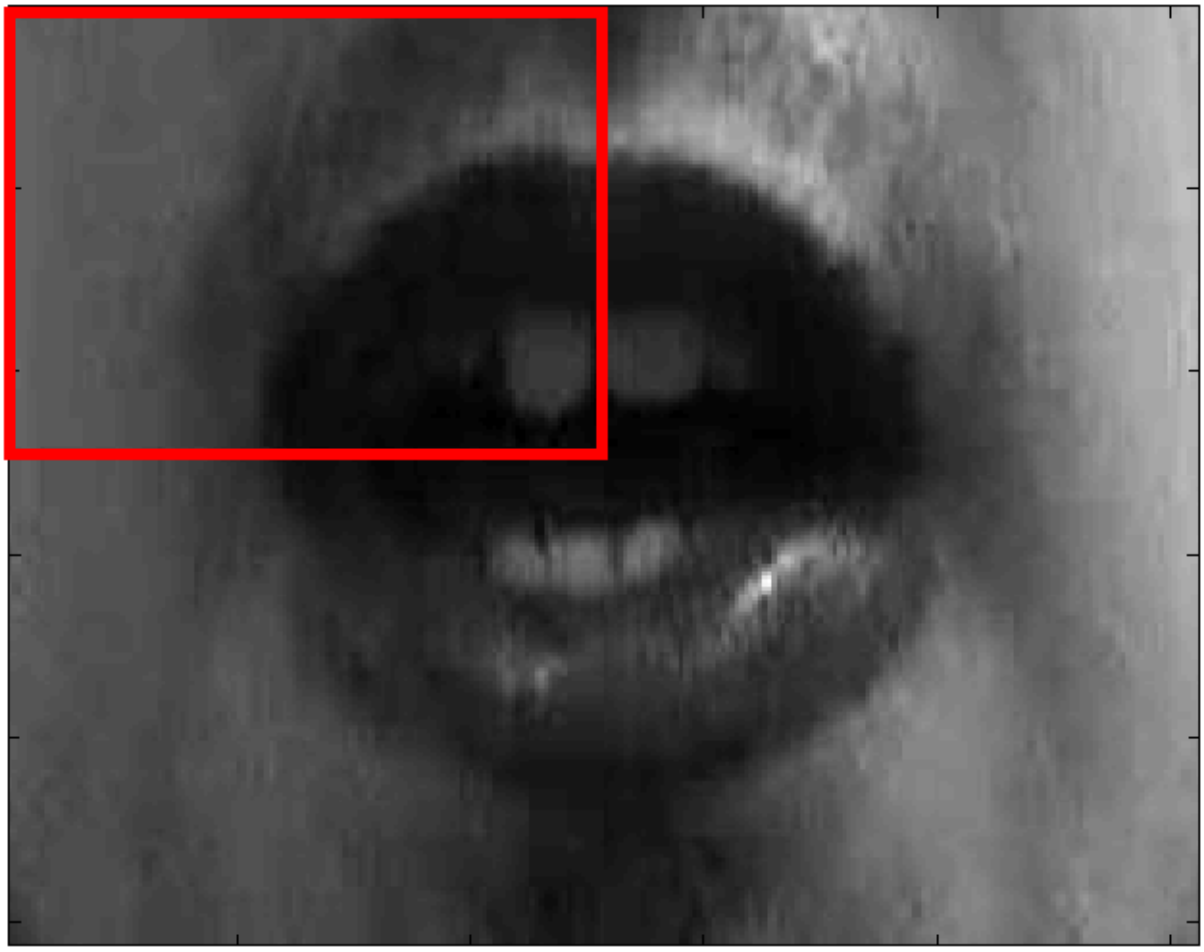

Figure 3 Example of frontal patch of a mouth. Example of the definition of a frontal patch of a mouth image necessary to the LLR computation.

case, LLR approximates the existing non-linear mapping between frequency features X_F and Y_F by distinct linear functions between the different frequency bands of the images.

Another option to apply pose normalization, is to project the final features X_L, Y_L used in the pattern classifier. Those features are obtained from linear dimensionality reduction transforms aimed at speech classification [3]. The transforms are usually based on LDA, which is a supervised transform projecting the DCT features to the linear subspace maximizing the separability of the C speech classes. Specifically, LDA finds the K-dimensional linear subspace maximizing the projected ratio $R = S_w^{-1} S_b$ between the inter-class scatter matrix S_b and intra-class scatter matrix S_w, defined as

$$S_w = \sum_{i=1}^{C} p_i \Sigma_i \qquad S_b = \sum_{i=1}^{C} p_i (\mu - \mu_i)(\mu - \mu_i)^T, \tag{4}$$

where p_i is the percentage of samples on the training set belonging to the class i, μ_i and Σ_i are the mean and covariance matrix for those samples and μ is the mean value of all the training samples in the dataset. The LDA projection matrix is then defined by the eigenvectors of R with K largest associated eigenvalues. If there is a linear mapping between the original features $x = Wy$, we can also relate the corresponding LDA projections with a linear mapping. Observing that

$$S_b^x = W S_b^u W^T \qquad S_w^x = W S_w^y W^T \tag{5}$$

it is easy to prove that if v is an eigenvector of R^y with eigenvalue λ_v, then $W^{-1}v$ is an eigenvector of R^x with the same eigenvalue and, consequently, there is also a linear mapping between the LDA projections associated to the frontal and lateral views. Two extra considerations have to be taken into account for the projection of the X_L and Y_L features. First, X_L and Y_L are obtained by applying LDA into the reduced DCT features X_F and Y_F,

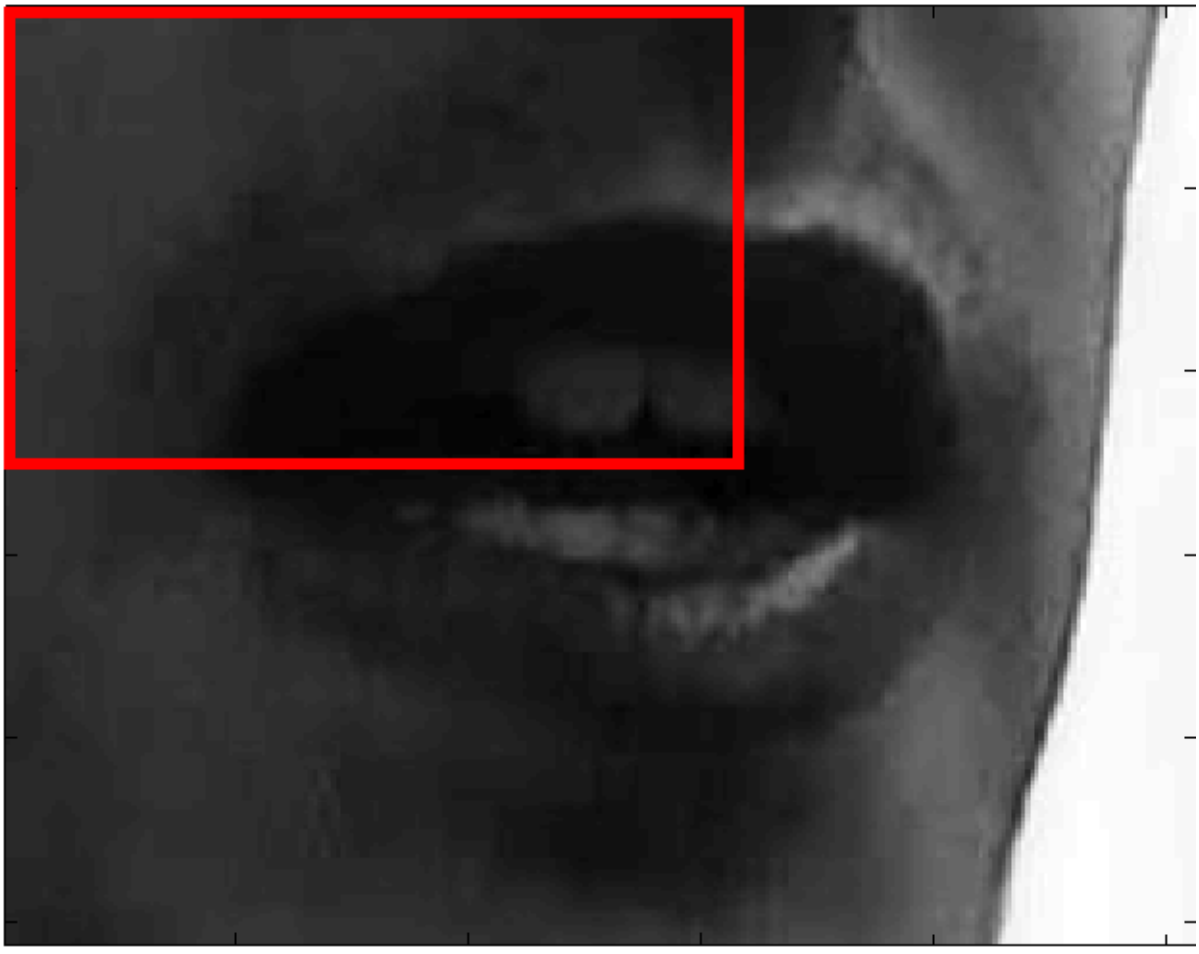

Figure 4 Example of lateral patch of a mouth. Example of the definition of a frontal patch of a mouth image necessary to the LLR computation.

which means that the projection by W^L is only a linear approximation of the real mapping between the LDA features in the same way W^F linearly approximates the relation between X_F and Y_F. Second, two stages of LDA are necessary to obtain X_L and Y_L from X_F and Y_F: first intra-frame LDA on the DCT features and then an inter-frame LDA on concatenated adjacent vectors extracted from the intra-frame LDA. In the intra-frame LDA, $x = x_F$, $y = y_F$, and $W = W^F$ in Equation (5), from which we obtain LDA projected vectors x_l and y_l, related with an approximated linear mapping W^l. In the inter-frame LDA, each x and y corresponds to the concatenation of 5 time-adjacent vectors x_l and y_l, and thus the approximated linear mapping W is given by a block matrix whose diagonal entries correspond to 5 block matrices W^l. As a consequence, if the linear approximation of $X_F = W^F Y_F$ holds, then it is also a valid assumption for the projection of the speech features by $X_L = W^L Y_L$.

The performance of the linear regression applied to the images or the extracted features can be analyzed by Equation (2) as the cost function Q normalized to the size of the vectors X and Y. The mean value taken by the cost function in our training dataset is presented in Figure 5. The curse of dimensionality is observed in the larger values of the mean square error in Q associated to the estimation of W^I in comparison to W^F or W^L. As experiments will show, the smaller dimensionality of the DCT and LDA features allow us to learn more accurately the GLR transform matrices, leading also to better speech recognition performance.

Consideration should be given to the fact that applying the pose normalization on the original images, or even to the low-frequency DCT coefficients, is independent of the features we posteriorly use for speech recognition and could be adopted with other appearance or contour-based visual speech features. The use of the LDA features, however, is specific to the speech

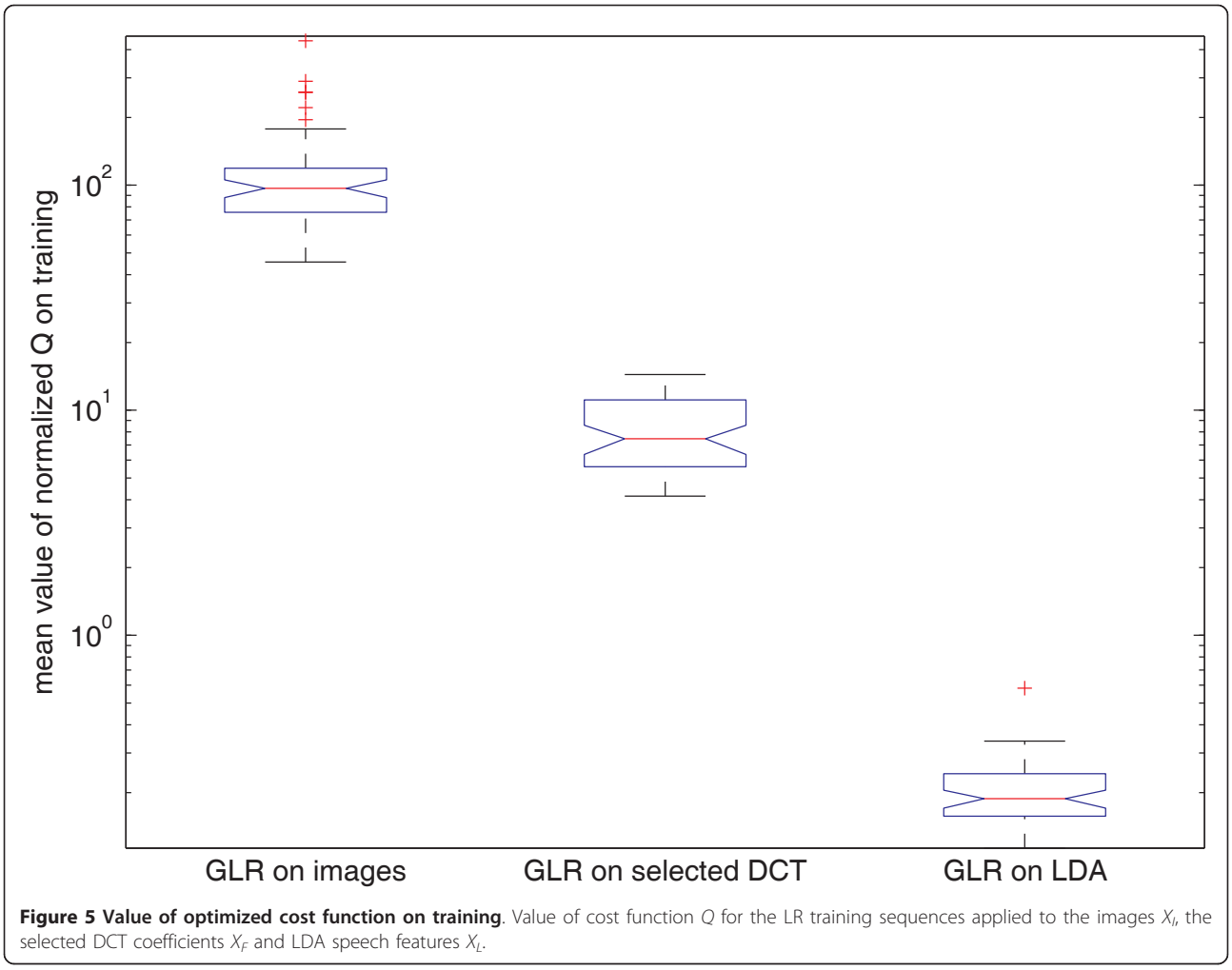

Figure 5 Value of optimized cost function on training. Value of cost function Q for the LR training sequences applied to the images X_I, the selected DCT coefficients X_F and LDA speech features X_L.

recognition system and involves an additional training of LDA projections for the different poses. In that sense, applying the LR techniques to the original images provides a more general strategy for the multi-pose problem, while the LDA features might be able to exploit their specificity for the speech recognition task.

3.4 Projective transforms on the images

A simple option when working with the images themselves is to estimate a projective transform from the lateral to the frontal views as a change of the coordinate systems between the images. In fact, as the difference in pose involves an extra dimension not taken into account in the projective model (3-D nature of the head rotation), that approach can only be justified for small pose changes, e.g., being impossible to implement for 90° of head rotation. We include in our experiments two projective transforms to measure the gain obtained by the learning approach of the LR techniques in comparison to projective transforms. In that case, we estimate a 3 × 3 projective transform T between the image coordinates in a semi-manual and automatic procedure. The coordinate points P used for that purpose are the corners of the lips, the center of the cupid's bow and the center of the lower lip contour for the different poses. In the manual procedure, we selected several frames of each sequence, manually marked the position of those four points for the frontal and lateral views and estimated the transform T minimizing the error of $P_{\text{frontal}} = TP_{\text{lateral}}$ over the selected frames of the sequence. For the automatic method, we segment the image based on color and region information and detect the lip's contour and the position of the points P from the segmentation. Examples of the images obtained with that method are shown in Figures 6, 7, and 8, where the deformations caused by neglecting the 3-D nature of head rotation are obvious. That effect is not encountered with the LR technique applied to the images, as the training stage on the process is responsible of learning how the mouth views change with the poses. In that

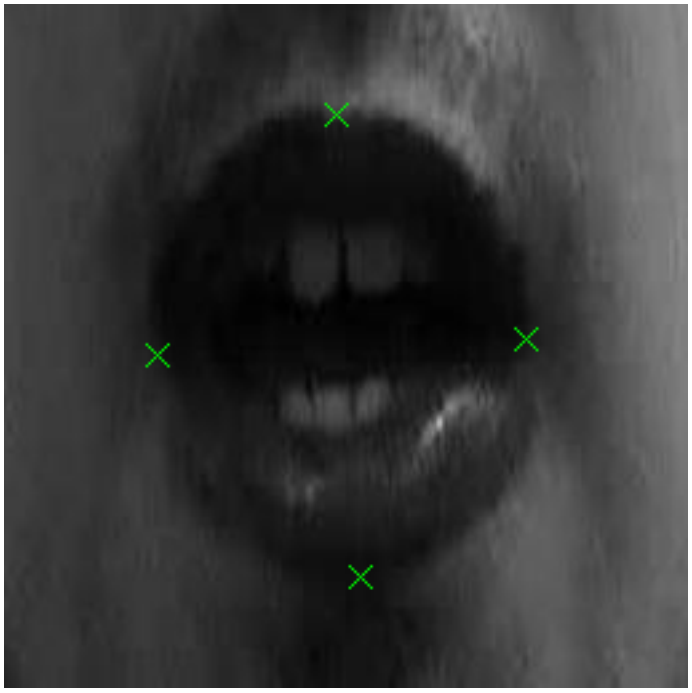

Figure 6 Manual annotation of frontal image. Original frontal image with manually annotated points P_{frontal} used in the estimation of a projective transform between images from different poses.

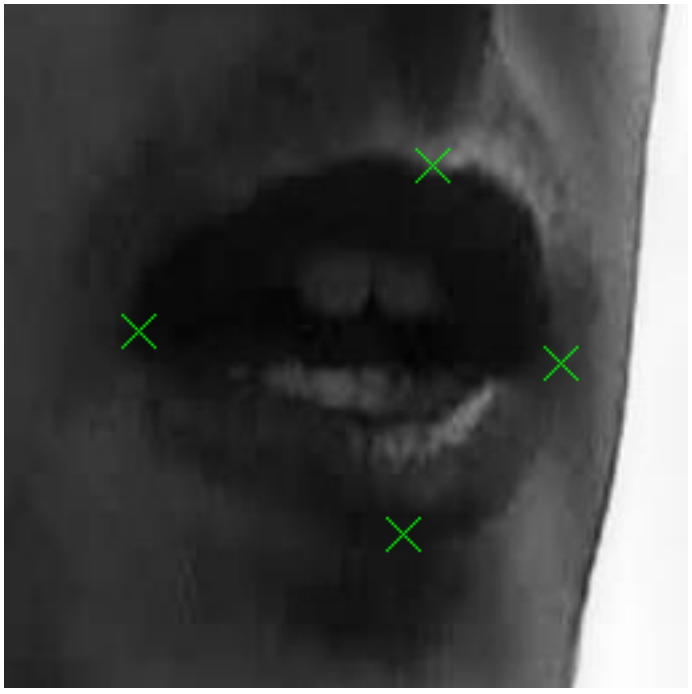

Figure 7 Manual annotation of frontal image. Original lateral image with manually annotated points P_{lateral} used in the estimation of a projective transform between images from different poses.

Figure 8 Virtual frontal image obtained by the projective transform. Virtual frontal image obtained by the projective transform T associated to the solution of $P_{\text{frontal}} = TP_{\text{lateral}}$ from Figures 6 and 7.

sense, the projective techniques can be used with any kind of images and do no exploit the fact that all the images correspond to mouths.

4 Experimental results

We present two sets of experiments: one on lipreading studying the adaptation of the visual stream to multi-pose conditions and another on AV-ASR analyzing the effects of the pose normalization on the audio-visual integration strategy. In lipreading experiments we first quantify the loss of performance associated to non-frontal poses, we then justify quantitatively the necessity of a pose normalization step and final analyze the performance of the proposed pose normalization strategies. In audio-visual experiments, we first study if the visual stream can still be exploited for speech recognition after the pose normalization has taken place, something that previous studies [4,5] on AV-ASR in realistic working conditions failed to do. In AV-ASR we are also interested in the influence of the pose normalization in the final performance and, specially, on the optimal value of the weight associated to the visual stream.

The technical details of the experimental set-up are the following. The task considered is connected speech recognition under different speaker poses relative to the camera. Training and testing has been done with the multi-speaker paradigm (all speakers are on train and test set but with different sequences) with three fold cross-validation and the results are given in terms of word accuracy. The same multi-speaker cross-validation is used to estimate the LR transforms for the different poses and features. The parameters of the feature extraction blocks and classifiers are chosen based on experiments with an evaluation dataset to optimize speech recognition. To fairly analyze the performance associated to frontal and lateral views, the same kind of classifiers are trained for each possible pose: frontal and lateral at 30°, 60° and 90° of head rotation. The HTK tool-kit [51] is used to implement three-state phoneme HMMs with a mixture of three Gaussians per state. For the multi-stream HMMs, the same number of states and Gaussians than in single-stream case is used. The parameters of the model are initialized with the values estimated for independent audio and visual HMMs and posteriorly re-estimated jointly. The audio and visual weights are considered fixed parameters of the system, restricted to sum up to one and optimized for speech recognition on the evaluation dataset.

4.1 Database

For our experiments, we required speech recordings with constrained non-ideal visual conditions, namely, fixed known poses and natural lighting. To that purpose we recorded our own database, which is publicly available at our webpage. It consists of recordings of 20 native French speakers with simultaneous different

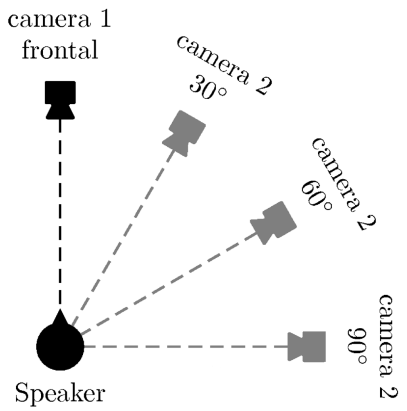

Figure 9 Schema of the simultaneous recordings. Schema of simultaneous recordings with different poses of the speaker. The first camera captures a frontal view of the speaker, while the second one records simultaneously a lateral view at 30°, 60°, or 90° of head rotation.

views, one always frontal to the speaker and the other with different lateral poses.

The recordings involve one frontal camera plus one camera rotated 30°, 60°, and 90° relative to the speaker in order to obtain two simultaneous views of each sequence, see Figures 9, 10, and 11. The first camera was fixed with a frontal view, while the second camera provided different lateral views. For each possible position of the second camera, the speaker repeated three times the digits, giving a total of three couples of repetitions of each digit for each pose: 9 for frontal views in total and 3 lateral repetitions for each possible degree of head rotation.

To comply with the natural conditions, the corpus was recorded with natural lighting conditions, resulting in shadows on some images under the nose and mouth of the subjects. The videos were recorded with two high-definition cameras CANON VIXIA HG20, providing 1920 × 1080 pixels resolution at 25 frames per second, and included the head and shoulders of the speaker.

Figure 10 Frontal view of one speaker from our database. Frontal view of one speaker from our database captured with the first camera.

Figure 11 Lateral view of one speaker from our database. Lateral view of one speaker from our database captured with the second camera at 60°. The image is the lateral view associated to the frontal image of Figure 10.

In terms of audio set-up, two different micros were used for the recordings, an external micro close to the speaker's mouth, without occluding its view, and the built-in micro of the second camera. That set-up provided two conditions for the audio signal, a clean audio signal obtained with an external microphone tailored for human voice and a noisy signal recorded with a lower quality microphone at some meters of distance to the speaker. Audio was recorded with a sample rate of 48000 Hz and 256 kbps for both micros and used to synchronize the videos, as it offered better time resolution than pairing of the video frames (offering only 40 milliseconds of frame resolution). For the two audio signals we computed the correlation of their normalized MFCC features within each manually segmented word, obtained an estimate of the a delay for each word and averaged over the whole sequence. The same delay was considered for the video signals, after correcting for the difference in distance between the two micros and the speaker.

The word labeling of the sequences was done manually at the millisecond and phone labels were posteriorly obtained by force alignment of the clean audio signals with the known transcriptions.

4.2 Visual speech recognition

In a first set of experiments we quantify the loss of performance of a lateral system compared to a fully frontal one. To that purpose, we paired the frontal and lateral sequences and test each sequence with the corresponding classifier, i.e., frontal sequences with frontal classifier and, for each possible head rotation, lateral sequences with their lateral classifier. That gives us a measure of how visual speech degrades with the different poses, presented in Figure 12. As happens with human lipreading [52], speech recognition deteriorates with non-frontal poses, which of course is more acute for 90° of head rotation (9% of loss of performance with respect to the frontal system) than for 60° (5% of loss of performance with respect to the frontal system). Figure 12 also shows

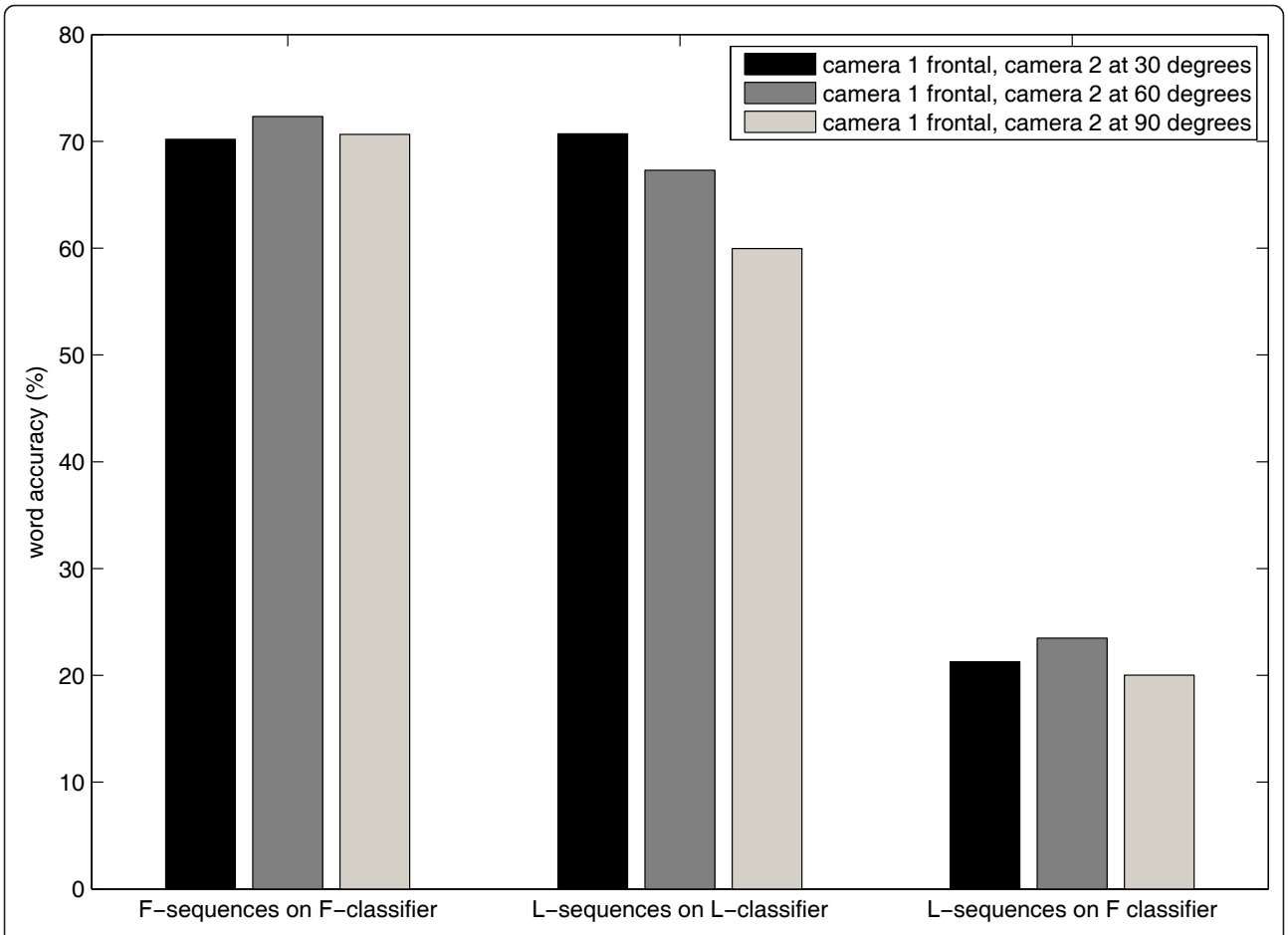

Figure 12 Lipreading performance associated to frontal and lateral views. This figure shows the loss of performance in visual-only experiments associated to non-frontal poses and the existence of a train/test pose mismatch when no pose normalization is applied to the visual stream. The first column corresponds to the frontal (F) sequences tested on the frontal classifier. The second column to the lateral (L) sequences with the corresponding lateral classifier. The last column to the lateral sequences directly tested on the frontal classifier, without the proposed pose normalization.

the performance of the frontal classifier tested with the lateral sequences when no pose normalization is applied, i.e., there is a mismatch on the train/test conditions in terms of pose and the system performs poorly, with mean word accuracy dropping from 71% to 22%. This justifies the necessity of pose normalization.

Next, we test the different pose normalization techniques on the lateral sequences with the classifier trained and optimized for frontal sequences. Figure 13 compares the results of the pose normalized lateral sequences to the corresponding frontal sequences with the frontal classifier and to the original lateral sequences on their lateral classifier. The results of the ideal frontal system represent the best we can do in terms of original pose and trained system, while the results of the completely lateral system represent the best we can do when the original images present a non-frontal pose with a lipreading system adapted to it. For each possible feature space, we choose the best-performing LR technique: LLR on the images and GLR on the selected DCT and LDA feature spaces. As expected, the features obtained after the pose normalization can neither beat the frontal system, because there is a loss of valuable information in the non-frontal images, nor obtain the performance of the corresponding lateral system, due to the limitations of the pose normalization techniques. For the different poses, the projected LDA features clearly outperform the other techniques (between 4% to 12% of loss of accuracy for the different poses compared to the frontal views) because they exploit the specificity of the features for speech recognition compared to the more general image or DCT feature spaces (accuracy loss 25% to 34% compared to the frontal views). As expected, the original images and the selected DCT coefficients present similar performance with different LR techniques and regularization parameters β. The accuracy of the LR

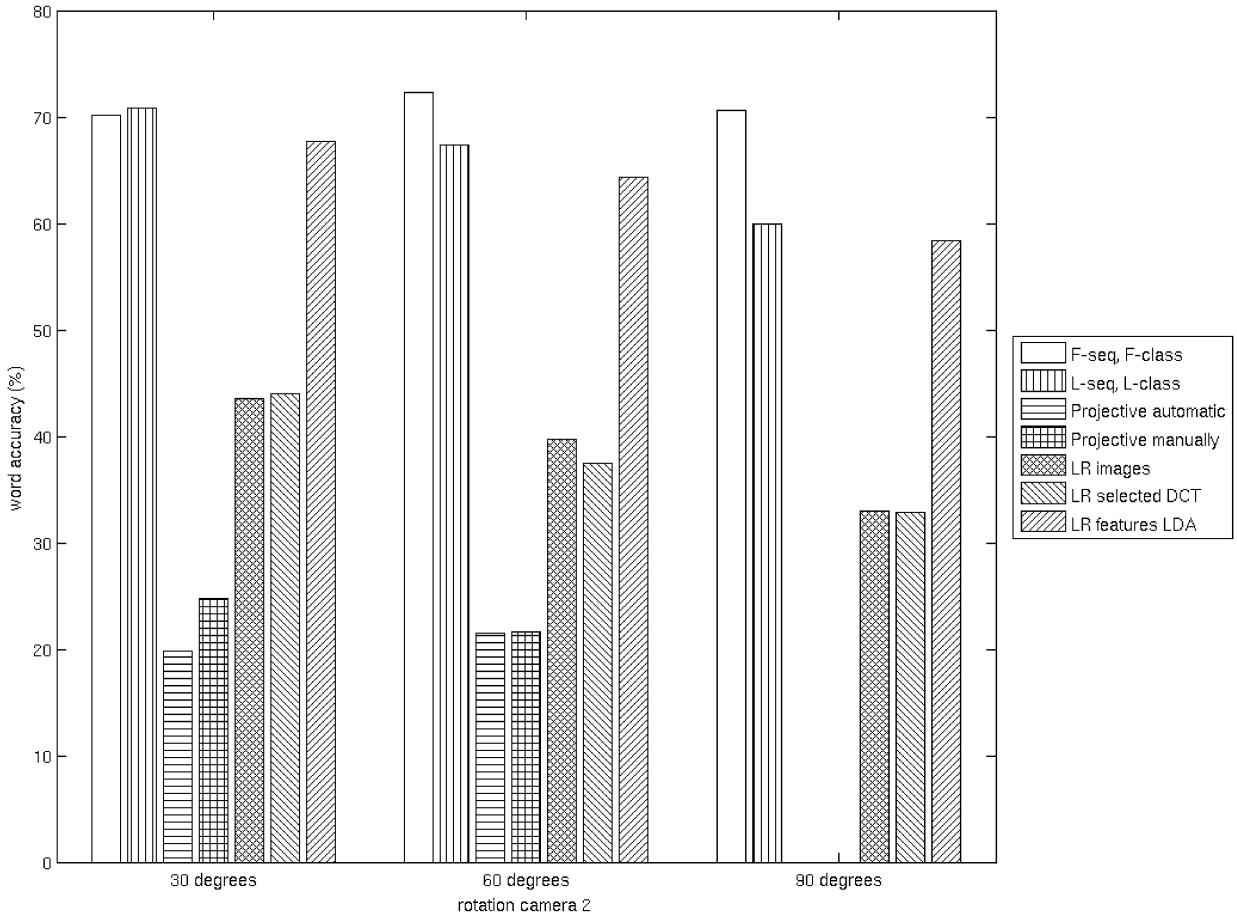

Figure 13 Lipreading with frontal, lateral and posed normalized sequences. The first column corresponds to the frontal sequences tested on the frontal classifier. The second column to the lateral sequences with the corresponding lateral classifier and the rest of columns to the pose-normalized lateral sequences tested on the frontal classifier. Third and forth columns corresponds to the projective transforms, while the last three correspond to the best LR technique applied to each feature space, i.e., LLR for the image space (with 32 × 32 pixel patches, 75% overlapping, β = 15) and GLR on the selected DCT (β = 5) and LDA (β = 0) feature spaces.

estimates decreases with the dimension of the feature space and, therefore, for the raw images it is necessary to both partition the image into patches to decrease the dimensionality of the LR estimates and to increase the value of the regularization parameter β. It is not worth then to work in the high dimensional image space with the LLR transform instead of applying the GLR to its reduced DCT features. Any improvement on the virtual views obtained in the LLR projection of images is lost on their posterior projection to the DCT space. Similarly, the projective transforms obtain poor recognition results (50% of accuracy loss compared to the frontal views) because they neglect the effects of 3-D pose changes on the views of the mouth.

Figure 14 compares the different LR techniques applied to the original images, where LLR performs better than GLR. Splitting the images in four patches and allowing an overlapping of 75% of the patches lead to the best results, showing the expected trade-off between the size-of the patch and the overlapping. A patch size too large (GLR case) suffers from the linear assumption and leads to blurring of the images, while a patch too small is more sensible to misalignments. Similarly, for each patch size, a high overlapping of patches results in over-smoothing, while low values cause block effects on the reconstructed images. At the same time, the value of the optimal regularization parameter β increases with the size of the patches.

For the selected DCT coefficients, however, the general mapping defined by GLR obtains better results, while for the LDA case both techniques perform similarly, see Figures 15 and 16. The worst performance of the LLR with the selected DCT features can be explained by the observation that the patches defined in

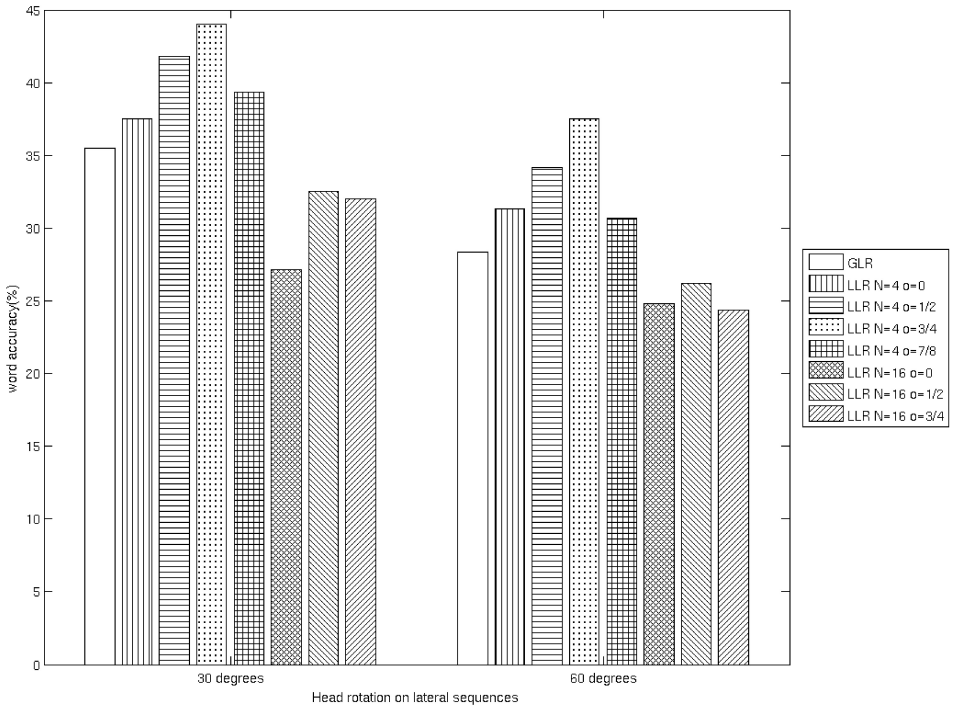

Figure 14 Lipreading performance with pose normalization in the image space. This figure shows the performance of a frontal classifier when the input corresponds to a lateral view of the speaker and the visual features have been normalized to a frontal pose in the image space. The experiments show the expected trade-off between the size-of the patch and the overlapping. N stands for the number of patches and o for the patch overlapping used in the LLR transform.

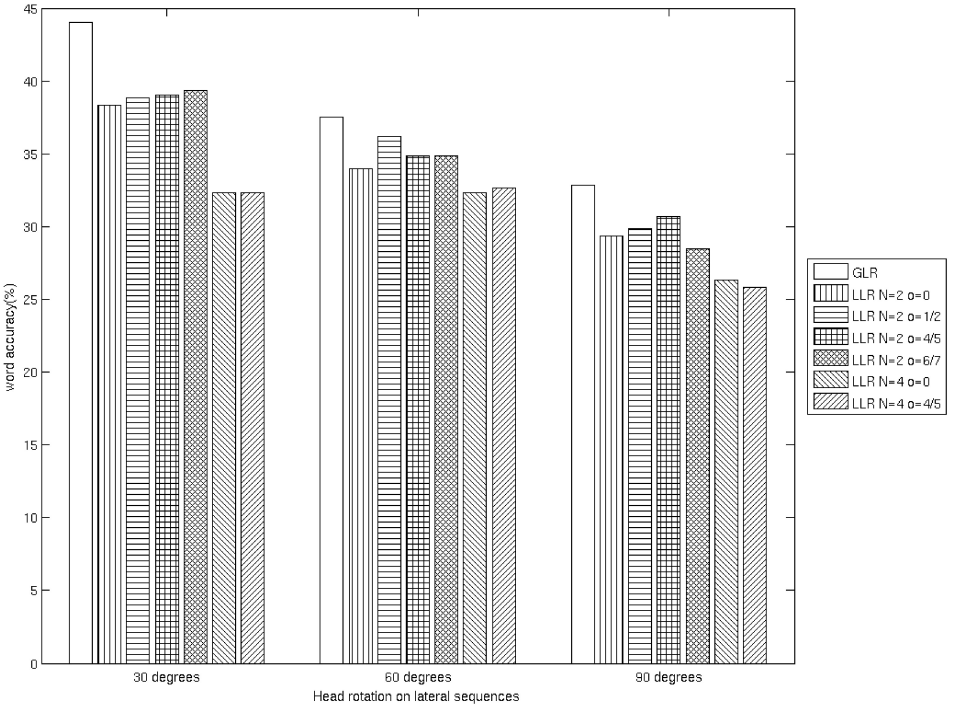

Figure 15 Lipreading performance with pose normalization in the DCT space. This figure shows the performance of a frontal classifier when the input corresponds to a lateral view of the speaker and the visual features have been normalized to a frontal pose in the DCT space. GLR outperforms LLR because the patches corresponding to high frequencies (containing the details of the image) can not be matched by a linear mapping.

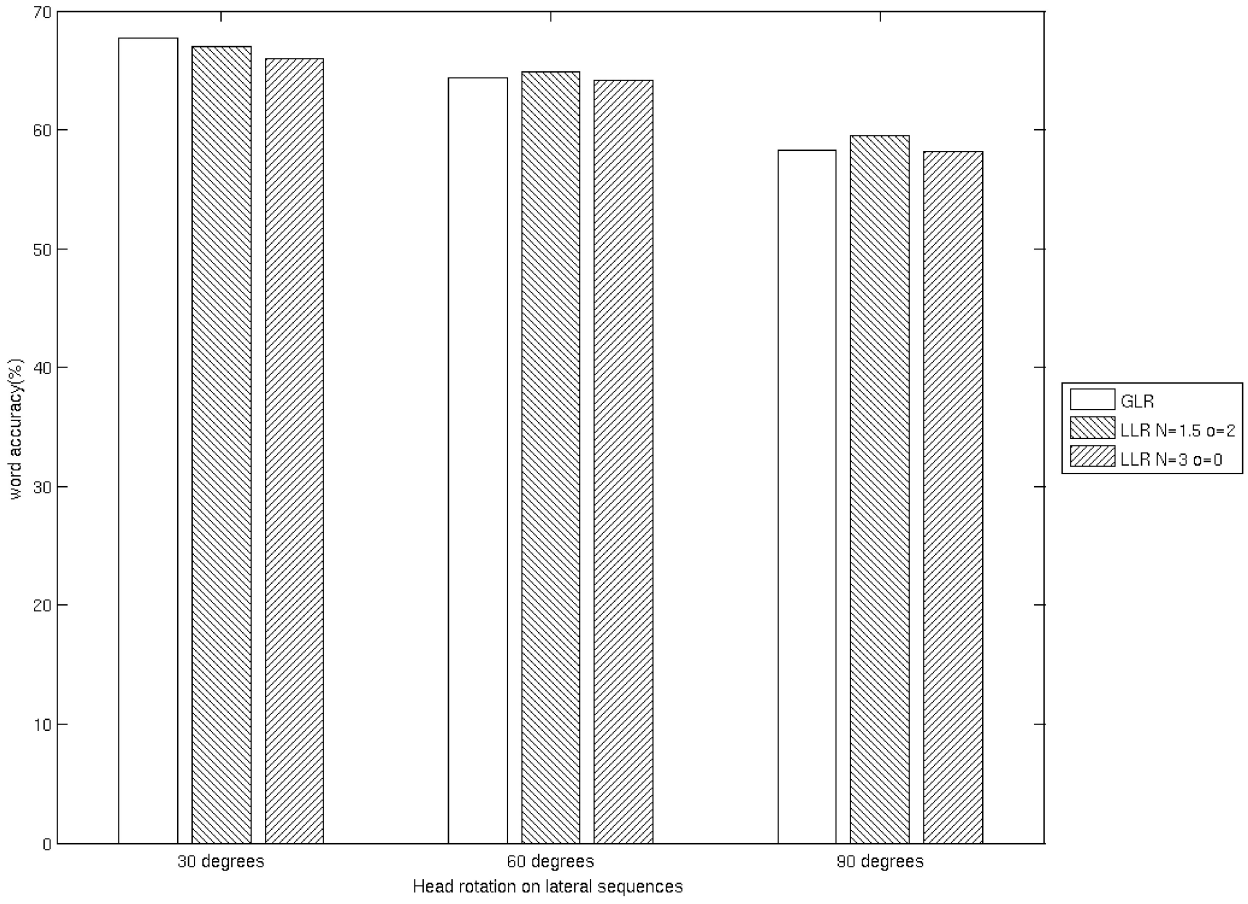

Figure 16 Lipreading performance with pose normalization in the LDA feature space. This figure shows the performance of a frontal classifier when the input corresponds to a lateral view of the speaker and the visual features have been normalized to a frontal pose in the LDA feature space. GLR and LLR perform similarly.

the DCT space correspond to high and low-frequency components of the images. It seems likely, therefore, that a linear transform between the low-frequency components of the images exist, but that assumption does not hold for the high-frequency components associated to image details. In the case of LDA features there is no interpretation of the patches defined on the LLR technique[a] and we observe that the mapping between the frontal and lateral features can be similarly approximated with the GLR and LLR techniques. In fact, there is no statistical difference between the performance of the GLR and LLR techniques in that feature space.

4.3 Audio-visual speech recognition

This set of experiments study how pose changes and normalization affects AV-ASR systems. Since the visual stream is most useful when the audio signal is corrupted, we report audio-visual experiments with a noisy audio signal and compare it to an audio-only ASR system. In an audio-visual system, the weight assigned to the visual stream controls to which extend the classifier's decision is based on the visual features, therefore differences between visual streams are more evident when the weight assigned to the video is high. The extreme cases correspond to a completely corrupted visual stream, where $\lambda_A = 1$, $\lambda_V = 0$ and the different pose normalization techniques obtain the same performance, and to a corrupted audio signal with weights $\lambda_A = 0$, $\lambda_V = 1$ and the lipreading performance already observed. Consequently, the differences in performance of the pose normalization methods are more acute with 0 dB than 7 dB of audio SNR and almost imperceptible with clean audio data. To that purpose we artificially added babble noise extracted from the NOISEX [53] database to the clean audio signal with 7 dB and 0 dB of SNR and test our pose normalization techniques in these conditions. The HMM audio parameters were trained in clean conditions, but tested with the

corrupted audio stream. The visual counter-part corresponds to the previous lipreading system, with the best GLR or LLR technique for each feature space.

Figures 17 and 18 show the performance for the audio-visual system for frontal and lateral poses. The performance of the different streams is coherent with the visual-only experiments, with frontal views outperforming lateral ones and GLR on the LDA space clearly outperforming the other pose normalization methods. The LR projection techniques applied to the original images or DCT coefficients are only able to improve audio recognition when the audio signal is highly corrupted (0 dB), while the projection on the LDA space always ameliorates the recognition of the audio system. The LR results for the images and DCT coefficients at 7 dB point out the fact those techniques are not useful for AV-ASR and, in fact, are outperformed by an audio-only system. In that case, the audio-visual system can not exploit the pose-normalized visual features because the errors incurred in the audio domain are not uncorrelated with the errors in the visual domain.

It is interesting to analyze the relation between the value of the optimal video weight λ_V assigned to the different pose normalization techniques and their performance in lipreading experiments. Figure 19 shows how the weight given to the visual modality decreases with the quality associated to the visual stream: for frontal sequences λ_V takes higher values than for the lateral ones, the projected lateral sequences have higher weights than the pose normalized ones, and the values for 90° of head rotation are lower than for 30°. Figure 20 shows that there is a clear correlation between the values of the optimal visual weight and the stream's performance in lipreading experiments. We can then conclude that the performance of the pose normalization techniques in lipreading is directly related to their

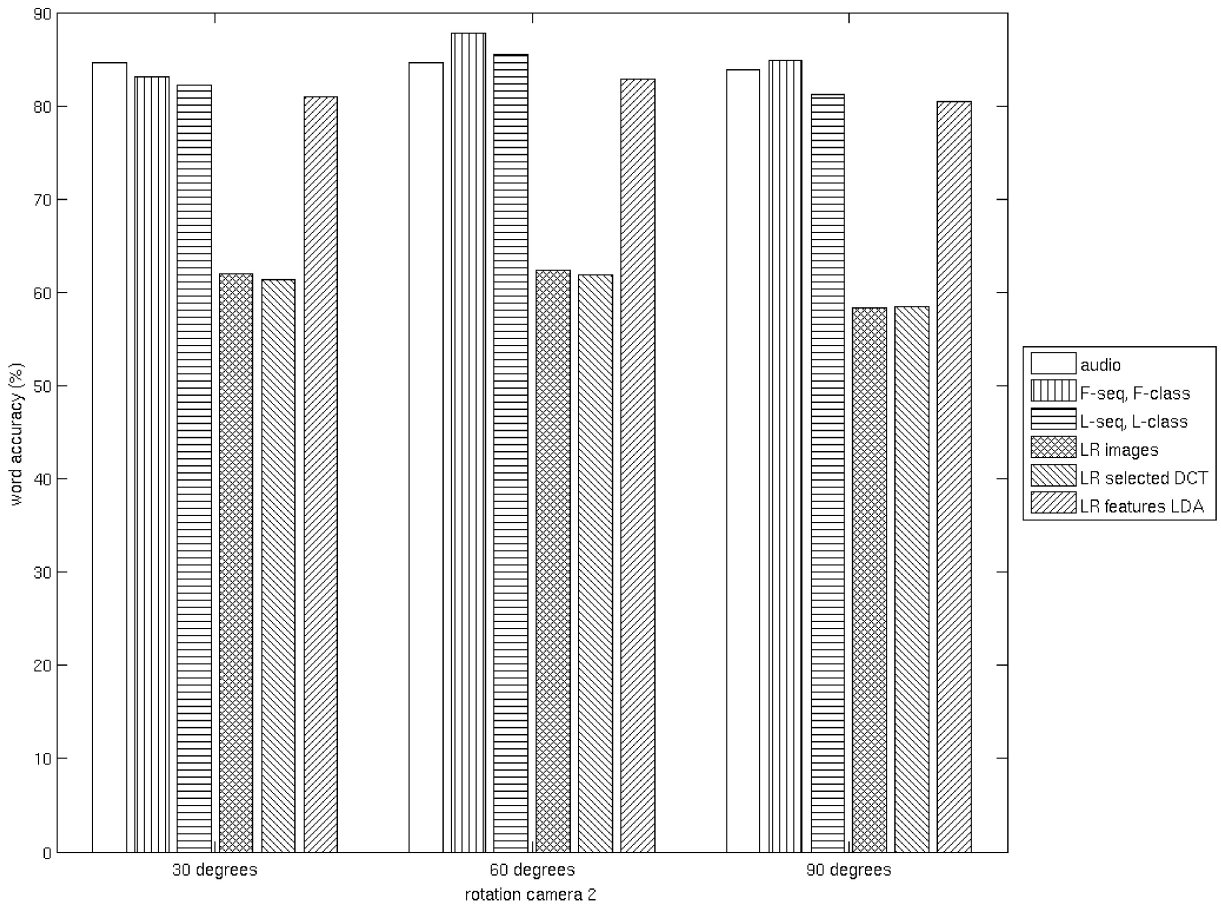

Figure 17 Performance of different AV-ASR systems with a corrupted audio stream with 7 dB of SNR. Mean word accuracy for audio and audio-visual systems with different visual streams and classifiers. The audio stream is a corrupted with 7 dB of SNR, while different visual streams are considered: the ideal frontal (F) views of the speaker, the original lateral (L) views at 30°, 60° and 90° of head rotation and the corresponding streams after pose normalization.

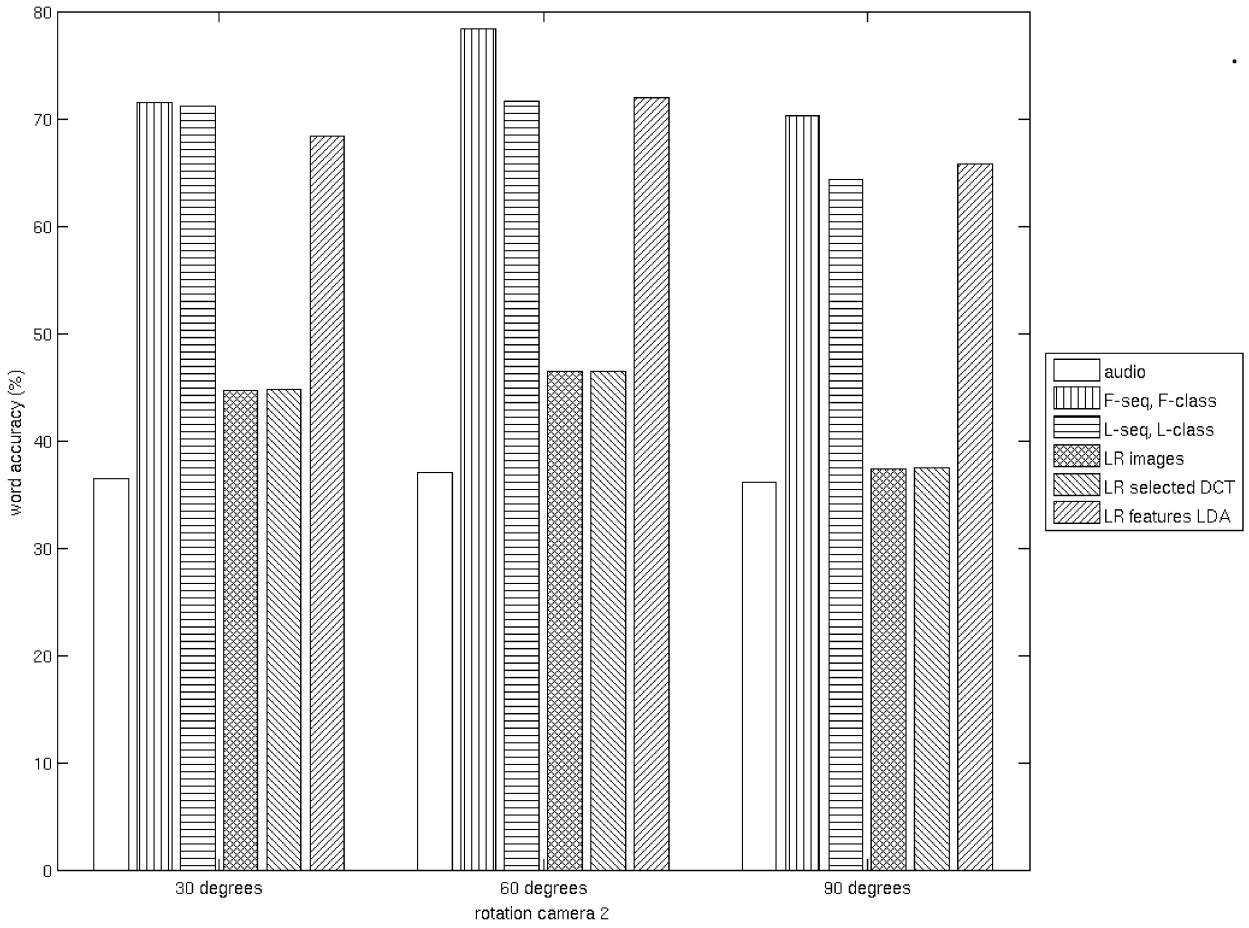

Figure 18 Performance of different AV-ASR systems with a corrupted audio stream with 0 dB of SNR. Mean word accuracy for audio and audio-visual systems with different visual streams and classifiers. The audio stream is a corrupted with 0 dB of SNR, while different visual streams are considered: the ideal frontal (F) views of the speaker, the original lateral (L) views at 30°, 60° and 90° of head rotation and the corresponding streams after pose normalization.

performance in audio-visual experiments and the value of weight assigned to the visual stream in the classifier.

4.4 Statistical significance of the results

In our experiments, we compare different views from the speaker and pose normalization strategies learned and tested on the same data and the results, therefore, reflect differences between the views and pose normalization strategies rather than differences in the test datasets. In this case, the statistical significance of the results cannot be evaluated by means of confidence intervals associated to the performance of each method independently, but requires the comparison of the different methods in a one-to-one basis for the same sentences, speakers and train/test datasets. In this study, we use the "probability of error reduction" p_e presented in [54] to assess the differences in performance of the proposed weighting schemes. We refer the reader to the original article [54] for a detailed description of p_e, give only an intuitive definition and use it to assess if one method significantly outperforms another. Intuitively, the probability of error reduction p_e between two systems A and B measures the number of independent testing samples that favor system A over B while leaving the rest of the samples unchanged.

To asses if the differences in performance between pose normalization applied in different feature spaces are statistically significant, we compute p_e with respect to a lateral system in the lipreading experiment. For the image and DCT feature spaces, performance degrades in every single test case for all the possible lateral views (p_e = 1). In the case of LDA feature space (with the GLR technique), performance degrades in 70% of the cases for 30° of head rotation and in 80% for the rest of the lateral views. We conclude that LR pose normalization is more successful in the LDA space, while the DCT

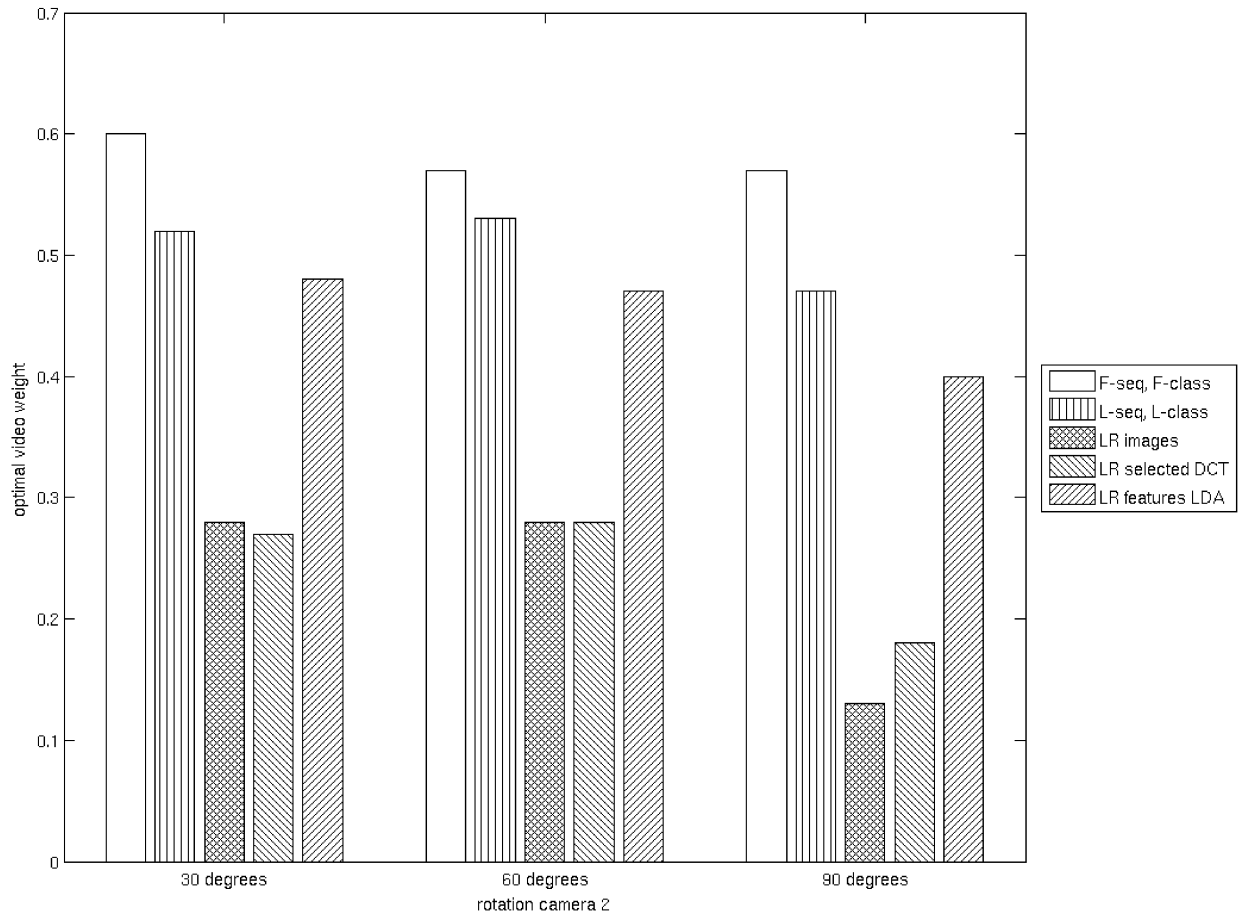

Figure 19 Optimal video weight in the multi-stream system with a corrupted audio stream with 7 dB of SNR. Optimal value of the weight associated to the visual stream in the audio-visual classifier. The audio stream is a corrupted with 7 dB of SNR, while different visual streams are considered: an ideal frontal view of the speaker, the original lateral views at 30°, 60° and 90° of head rotation and the corresponding streams after pose normalization.

and image spaces perform poorly. At the same time, even though the final accuracy of the lateral system is close to the projected LDA features, there is a significant loss of performance due to the pose normalization.

For the audio-visual experiments, we compare each of the systems to an audio-only recognizer. Only the pose normalization in the LDA space is able to exploit the visual stream with 7 dB of SNR, with performance improving in 98%, 95%, and 89% of the sequences at 30, 60, and 90° in comparison to an audio-only system. This percentage is inferior to 16% and 13% for the DCT or image space, pointing out that pose normalization in these feature spaces fails to exploit the visual modality in an AV-ASR system. In a more noisy environment with 0 dB of SNR, the projection on the LDA space is always beneficial, while the DCT and image spaces only do better than an audio-only system in 80% of the cases. This analysis confirms that pose normalization is only really successful in the LDA feature space in both visual and audio-visual ASR systems.

5 Conclusions

In this article, we presented a lipreading system able to recognize speech from different views of the speaker. Inspired by pose-invariant face recognition studies, we introduce a pose normalization block in a standard system and generate virtual frontal views from non-frontal images. In particular, we use linear regression to project the features associated to different poses at different stages of the lipreading system: the images themselves, a low-dimensional and compact representation of the images in the frequency domain or the final LDA features used for classification. Our experiments show that the pose normalization is more successful when applied directly to the LDA features used in the classifier, while the projection of more general features like the images

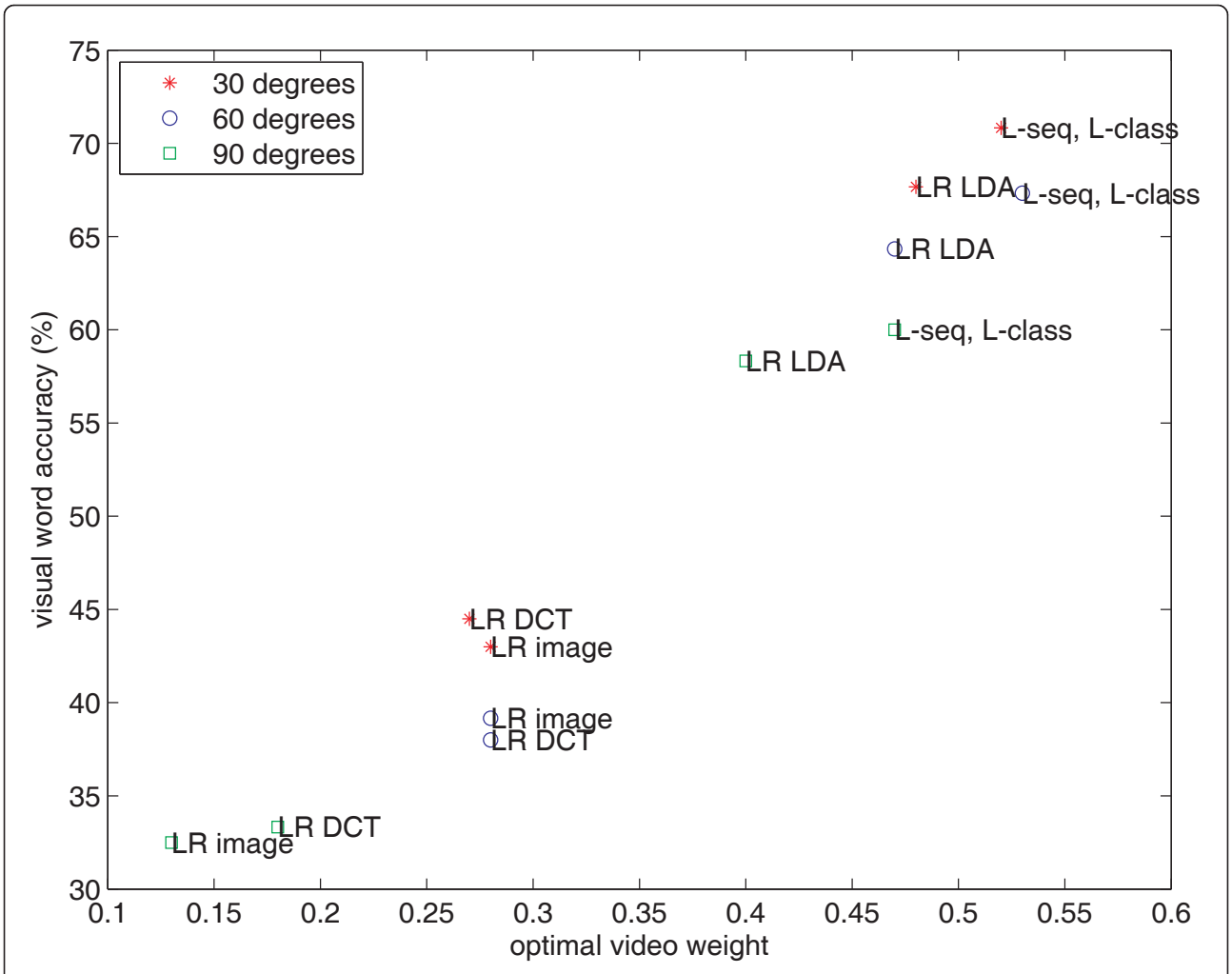

Figure 20 Scatter between optimal video weight and lipreading performance for a corrupted audio with 7 dB of SNR. Scatter plot between the optimal value of the weight associated to the visual stream in the audio-visual classifier and the performance of the corresponding visual stream in lipreading experiments. The audio stream is a corrupted with 7 dB of SNR, while different visual streams are considered: an ideal frontal view of the speaker, the original lateral views at 30°, 60°, and 90° of head rotation and the corresponding streams after pose normalization.

or their low-frequency representation fails because of misalignments on the training data and errors on the estimation of the transforms.

In terms of AV-ASR, we study the effects of pose normalization in the fusion strategy of the audio and visual modalities. We evaluate the effects of pose normalization on the weight associated to the visual stream and analyze for which one of the proposed techniques the audio-visual system is able to exploit its visual modality. We show that only the projection of the LDA features used in the classifier is really able to normalize the visual stream to a virtual frontal pose and enhance the performance of the audio system. As expected, there is a direct relation between the optimal weight associated to the pose normalized visual stream and its performance in lipreading experiments. Consequently, we can simply study the effects of pose normalization in the visual domain and transfer the improvements into the audio-visual task by adapting the weight associated to the visual stream.

Endnotes

[a]For simple LDA we can interpret the patches as directions on the original space maximizing the projected ratio R, so that if we sort the eigenvectors on the LDA projection according to their eigenvalue, we could interpret the patches as linear subspaces decreasingly maximizing the projected ratio. However, as we include intra and inter-frame LDA in the W^L transform, no interpretation is possible for the patch definition on the x_L, y_L space.

Acknowledgements
This study was supported by the Swiss SNF grant number 200021-130152.

Competing interests
The authors declare that they have no competing interests.

Received: 3 October 2011 Accepted: 29 February 2012
Published: 29 February 2012

References

1. G Potamianos, C Neti, J Luettin, I Matthews, Audio-visual automatic speech recognition: an overview, in *Issues in audio-visual speech processing*, ed. by Bailly G, Vatikiotis-Bateson E, Perrier P (MIT Press, Cambridge, 2004)
2. S Dupont, J Luettin, Audio-visual speech modeling for continuous speech recognition. IEEE Trans Multimedia. **2**, 141–151 (2000). doi:10.1109/6046.865479
3. G Potamianos, C Neti, G Gravier, A Garg, A Senior, Recent advances in the automatic recognition of audio-visual speech. Proc IEEE. **91**(9), 1306–1326 (2003). doi:10.1109/JPROC.2003.817150
4. G Potamianos, C Neti, Audio-visual speech recognition in challenging environments. in *Eighth European Conference on Speech Communication and Technology*, EUROSPEECH-2003 1293–1296 (2003)
5. K Livescu, O Cetin, M Hasegawa-Johnson, S King, C Bartels, N Borges, A Kantor, Lal P, L Yung, A Bezman, S Dawson-Haggerty, B Woods, J Frankel, M Magami-Doss, K Saenko, Articulatory feature-based methods for acoustic and audio-visual speech recognition. in *Final Workshop Report, Center for Language and Speech Processing, John Hopkins University*. **4**, 621–624 (2006)
6. S Furui, Cepstral analysis technique for automatic speaker verification. IEEE Trans Acoustics Speech Signal Process. **29**(2), 254–272 (2003)
7. H Hermansky, N Morgan, RASTA processing of speech. IEEE Trans Speech Audio Process. **2**(4), 578–589 (1994). doi:10.1109/89.326616
8. V Blanz, P Grother, P Phillips, T Vetter, Face recognition based on frontal views generated from non-frontal images. IEEE Proc Comput Vision Pattern Recogn. **2**, 454–461 (2005)
9. V Blanz, T Vetter, Face recognition based on fitting a 3D morphable model. IEEE Trans Pattern Anal Mach Intell. **25**(9), 1063–1074 (2003). doi:10.1109/TPAMI.2003.1227983
10. M Wai Lee, S Ranganath, Pose-invariant face recognition using a 3D deformable model. Pattern Recogn. **36**(8), 1835–1846 (2003). doi:10.1016/S0031-3203(03)00008-6
11. X Chai, S Shan, X Chen, W Gao, Locally linear regression for pose-invariant face recognition. IEEE Trans Image Process. **16**(7), 1716–1725 (2007)
12. X Zhang, CC Broun, RM Mersereau, MA Clements, Automatic speechreading with applications to human-computer interfaces. EURASIP J Appl Signal Process. **2002**, 1228–1247 (2002). doi:10.1155/S1110865702206137
13. P Lucey, G Potamianos, S Sridharan, A unified approach to multi-pose audio-visual ASR, in *8th Annual Conference of the International Speech Communication Association*, Interspeeech, (Antwerp, Belgium, 2007), pp. 650–653
14. P Lucey, G Potamianos, S Sridharan, An Extended Pose-Invariant Lipread-ing System, in *International Workshop on Auditory-Visual Speech Processing*, ed. by Vroomen, Jean, Swerts, Marc, Krahmer, Emiel (Hilvarenbeek, 2007)
15. V Estellers, JP Thiran, Multipose Audio-Visual Speech Recognition, in *19th European Signal Processing Conference EUSIPCO*, vol. 2011. (Barcelona, 2011), pp. 1065–1069
16. H Hermansky, Perceptual Linear Predictive (PLP) Analysis of Speech. J Acoustical Soc Am. **87**(4), 1738–1752 (1990). doi:10.1121/1.399423
17. P Mermelstein, Distance measures for speech recognition, psychological and instrumental. in *Pattern Recognition and Artificial Intelligence*. **116**, 374–388 (1976)
18. SB Davis, P Mermelstein, Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans Acoustics Speech Signal Process. **28**(4), 357–366 (1980). doi:10.1109/TASSP.1980.1163420
19. H Cetingul, Y Yemez, E Erzin, A Tekalp, Discriminative analysis of lip motion features for speaker identification and speech-reading. IEEE Trans Image Process. **15**(10), 2879–2891 (2006)
20. L Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition. Proc IEEE. **77**(2), 257–286 (1989). doi:10.1109/5.18626
21. C Bregler, Y Konig, Eigenlips for robust speech recognition. in *1994 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1994. ICASSP-94. **2**, II/669–II/672 (1994). vol. ii
22. M Tomlinson, M Russell, N Brooke, Integrating audio and visual information to provide highly robust speech recognition. in *Conference Proceedings., 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1996. ICASSP-96. **2**, 821–824 (1996)
23. M Kaynak, Zhi Q, A Cheok, K Sengupta, Z Jian, KC Chung, Analysis of lip geometric features for audio-visual speech recognition. Systems IEEE Trans Man Cybernetics, Part A: Syst Humans. **34**(4), 564–570 (2004). doi:10.1109/TSMCA.2004.826274
24. G Potamianos, HP Graf, E Cosatto, An image transform approach for HMM based automatic lipreading. in *IEEE International Conference on Image Processing*, Chicago. **II**, 173–177 (1998)
25. P Scanlon, D Ellis, R Reilly, Using mutual information to design class specific phone recognizers, in *Eighth European Conference on Speech Communication and Technology*, (EUROSPEECH-2003, 2003), pp. 857–860
26. EK Patterson, S Gurbuz, Z Tufekci, JN Gowdy, Moving-talker, speaker-independent feature study, and baseline results using the WAVE Multimodal speech corpus. Eurasip J Appl Signal Process. **2002**(11), 1189–1201 (2002). doi:10.1155/S1110865702206101
27. M Sonka, V Hlavac, R Boyle, Image Processing, Analysis and Machine Vision. International Thomson. **35**(1), 102–104 (1999)
28. P Lachenbruch, M Goldstein, Discriminant analysis. Biometrics. **35**, 69–85 (1979). doi:10.2307/2529937
29. R Battiti, Using mutual information for selecting features in supervised neural net learning. IEEE Trans Neural Netw. **5**(4), 537–550 (1994). doi:10.1109/72.298224
30. F Fleuret, Fast binary feature selection with conditional mutual information. J Mach Learn Res. **5**, 1531–1555 (2004)
31. H Peng, F Long, C Ding, Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell. **27**(8), 1226–1238 (2005)
32. M Gurban, JP Thiran, Information theoretic feature extraction for audiovisual speech recognition. IEEE Trans Signal Process. **57**(12), 4765–4776 (2009)
33. A Adjoudani, C Benoit, *Speechreading by Humans and Machines: Models, Systems and Applications*, (Springer, Berlin, 1996), pp. 461–471. 150 NATO ASI Series F
34. T Chen, Audiovisual speech processing. IEEE Signal Process Mag. **18**(1), 9–21 (2001). doi:10.1109/79.911195
35. C Neti, G Potamianos, J Luettin, I Matthews, H Glotin, D Vergyri, Large-vocabulary audio-visual speech recognition: a summary of the Johns Hopkins Summer 2000 Workshop, *Proc Works Signal Processing*, (Cannes, 2001), pp. 619–624
36. G Potamianos, J Luettin, C Neti, Hierarchical discriminant features for audio-visual LVCSR. *International Conference on Acoustics Speech and Signal Processing ICASSP*. **1**, 165–168 (2001)
37. J Movellan, G Chadderdon, Channel separability in the audio-visual integration of speech: A Bayesian approach. NATO ASI Series F Comput Syst Sci. **150**, 473–488 (1996)
38. D Massaro, D Stork, Speech recognition and sensory integration. Am Sci. **86**(3), 236–244 (1998)
39. J Kittler, M Hatef, R Duin, J Matas, On combining classifiers. IEEE Trans Pattern Anal Mach Intell. **20**(3), 226–239 (1998). doi:10.1109/34.667881
40. K Kirchhoff, J Bilmes, Dynamic classifier combination in hybrid speech recognition systems using utterance-level confidence values. in *International Conference on Acoustics, Speech, and Signal Processing*. **2**, 693–696 (1999)
41. L Rabiner, B Juang, An introduction to Hidden Markov models. IEEE ASSP Mag. **3**, 4–16 (1986)
42. C Bishop, *Neural Networks for Pattern Recognition*, (Oxford University Press, Oxford, 1995)
43. L Rabiner, BH Juang, *Fundamentals of Speech Recognition*, (Signal processing, Prentice Hall, NJ, 1993)
44. R Gross, I Matthews, S Baker, Appearance-based face recognition and light-fields. IEEE Trans Pattern Anal Mach Intell. **26**(4), 449–465 (2004). doi:10.1109/TPAMI.2004.1265861
45. V Blanz, T Vetter, Face recognition based on fitting a 3D morphable model. IEEE Trans Pattern Anal Mach Intell. **25**(9), 1063–1074 (2003). doi:10.1109/TPAMI.2003.1227983
46. T Vetter, Synthesis of novel views from a single face image. Int J Comput Vision. **28**(2), 103–116 (1998). doi:10.1023/A:1008058932445

47. D Beymer, Face recognition under varying pose. IEEE Proc Comput Vision Pattern Recogn. **1**, 756–761 (1994)
48. Q Summerfield, *Hearing by Eye: The Psychology of Lip-Reading*, (Lawrence Erlbaum Associates, Hillsdale, NJ, 1987)
49. CM Bishop, NM Nasrabadi, *Pattern Recognition and Machine Learning*, (Springer, New York, 2006)
50. R Bellman, in *Adaptive Control Processes: a guided tour*, vol. 1. (Princeton University Press, Princeton, 1961), p. 2
51. S Young, G Evermann, D Kershaw, G Moore, J Odell, D Ollason, V Valtchev, P Woodland, in *The HTK book*, vol. 2. (Cambridge University Engineering Department, 1997)
52. T Jordan, S Thomas, Effects of horizontal viewing angle on visual and audiovisual speech recognition. J Exp Psychol. **27**(6), 1386–1403 (2001)
53. A Varga, H Steeneken, M Tomlinson, D Jones, *The NOISEX-92 study on the Effect of Additive Noise on Automatic Speech Recognition*, (Tech. Rep., DRA Speech Research Unit, Malvern, England, 1992)
54. M Bisani, H Ney, Bootstrap estimates for confidence intervals in ASR performance evaluation. *International Conference on Acoustics, Speech, and Signal Processing*. **1**, 409–412 (2004)

doi:10.1186/1687-6180-2012-51
Cite this article as: Estellers and Thiran: **Multi-pose lipreading and audio-visual speech recognition.** *EURASIP Journal on Advances in Signal Processing* 2012 **2012**:51.